



BEYOND RECOGNITION

Technology now enables computers to identify you without you having to lift a finger, but will it protect your liberties – or undermine them? Jon Thompson considers how your movements are monitored with calm-computing analysis

Your smartphone is peeking enticingly out of your pocket. An opportunistic thief picks your pocket, but before he's got 10 yards away, your phone is shouting "stop thief", dialling 999 and sending out a location detector signal. It then switches itself off, so the thief cannot access its contents.

Given the number of expensive gadgets we carry around with us, it would be comforting if we could rely on safety measures such as these should the worst happen. Several research groups say it's possible to identify a person from the way they type and surf the web, and even the way they walk – consolation for anyone who has been mugged for their mobile or laptop or fallen prey to an online identity thief. There is a darker side to this technology, though. It could also be used to monitor your movements, no matter how carefully you tried to hide your identity. In this feature, we take an in-depth look at this type of technology, how close it is to reality and just how accurate it might be.

CODE COMFORT

Jani Mäntyjärvi and colleagues at Finland's VTT Electronics in Oulu have been worried about the security of modern portable gadgets. "These

devices contain increasing amounts of valuable personal information. For example, some smartphones contain wallet and e-commerce applications. Thus, the risk associated with them is increasing. Currently... sensitive data in mobile phones is protected with PIN codes. But in the 'on' state, not even the PIN code protects the information."

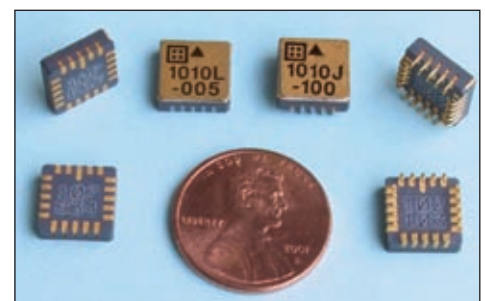
Leave your phone unlocked on a table in a café, go to pay the bill and, by the time you return, someone else could have access to your life. However, Mäntyjärvi believes he can make a phone identify who's carrying it. This idea uses some of the principles behind what's known as 'calm' computing. For an application to be considered calm, it mustn't demand action from the user. An application isn't considered calm when you need to take an action, such as entering a PIN or speaking a password for voice-recognition purposes.

GAIT EXPECTATIONS

A security system that recognises whether you're carrying it or not is an example of calm computing. This technology, called 'gait recognition', has long been used in computer vision systems. The VTT team wondered if

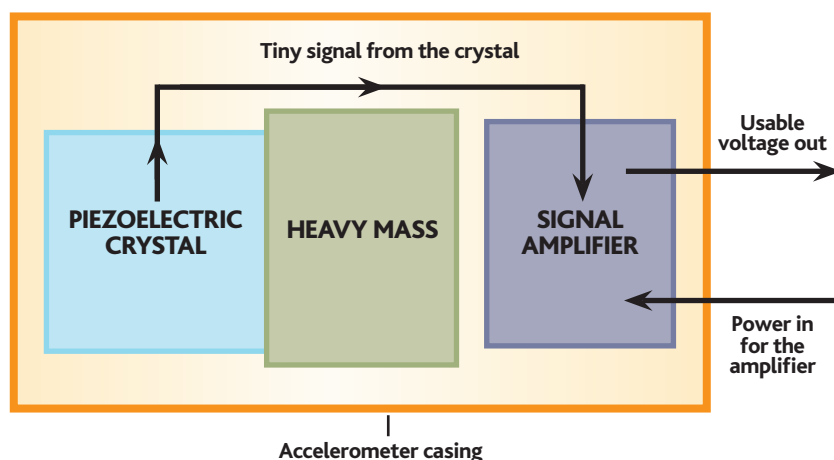
they could directly capture and use the natural rhythmic motion a person generates when they walk. In theory, how someone walks relies on unique factors such as height, weight, bone and muscle structure, overall fitness and so on. As you walk, the position of your body changes rhythmically in 3D space.

Are these rhythms unique enough to identify someone? To find out, the team took an accelerometer and mounted it on a belt. A group of 19 men and 17 women then took turns to wear the belt and walk at slow, normal and brisk paces while it digitised and recorded the data



▲ The tiny accelerometers pioneered by VTT can identify people just from the way they walk

Pressure drop How accelerometers work



▲ An accelerometer comprises three main components. When the unit is moved, a heavy mass presses against a tiny piezoelectric crystal. When you compress such a crystal, it generates a tiny electric charge that is proportional to the pressure applied. This is proportional to the acceleration the device is undergoing. The signal then passes to an onboard, externally powered signal amplifier, which creates a usable voltage. Place three accelerometers at right angles to each other and they can detect movement in 3D space. These tiny devices find uses in everything from detecting crashes in air bag systems to identifying probe touchdowns on other planets.

generated by the accelerometer. This was the team's reference point. If the experiment was to work, software running on an associated laptop should be able to analyse future data and compare it against this reference, indicating who was subsequently wearing the device (see 'Pressure drop', above). A week later, the volunteers were asked to wear the belt again and walk while it captured more data. Results showed that the software could identify who was doing the walking.

One key consideration was how long a person had to walk around until the software could get a fix on his or her identity. In this test, recognition speeds were fast. By taking 256 slices of data per second from the accelerometer, the team found that the software could identify individuals after just 20 metres of walking. So if someone tried to walk off with your phone, by the time they were around the corner the phone would know it wasn't you carrying it and either shut down or call for help – at least until you used a less 'calm' means of identification, such as keying in a PIN, for instance.

Mäntyjärvi sees uses for his discovery not only in smartphones but also in mobile banking and credit card payment applications. However, it will be some time before such protection becomes available, as there are still hurdles to overcome. "The potential drawbacks of the method are partly common to all gait-based methods," he warned. "Effect of changes in the speed of walking, [high heels] and ground; also drunkenness and injuries affect gait." Mäntyjärvi also said he needed to understand the effect of carrying the device on different parts of the body. The data gathered from a phone swinging along in a bag might be very different from that gathered when it's in a coat pocket.

INTERNET USE ANALYSIS

Other groups are investigating 'calm' methods of identifying us to the systems we use, such as websites. Detecting who's online from their unique usage patterns, for instance, might save the billions lost every year to online fraud, but

only if such signatures are clear enough to give a positive identification first time every time.

One group of researchers believes these signatures do exist, and that they know how to spot them. Researchers Balaji Padmanabhan and Yinghui Yang have come up with clickprints, which they say are analogous to fingerprints. Clickprints, they claim, are capable of spotting someone online in a mass of visitors after just a few visits. In a complex online paper, Padmanabhan, of the Wharton School at the University of Pennsylvania, and Yang, of the University of California at Davis, revealed an algorithm that takes as its inputs things such as the time of access, source IP address, referring page, number of webpages visited and the time spent on each page. The output is a prediction of who is doing the surfing.

To determine its accuracy, Padmanabhan and Yang tested it using website usage data from comScore Networks. In much the same way that market research companies gather TV viewing habits, comScore provides real-time measurements of internet use, all captured from a consenting panel of one million users worldwide. For their tests, the researchers used data for the surfing habits from 50,000 comScore users. They also selected five popular sites, which remain anonymous in their results. From these 50,000 visitors, they chose just five for an initial test. The researchers selected five features from the site usage data. These were the duration of a visit to a site, the number of pages viewed, the average time spent on each page, the time the user visited each page and the day of the week on which the visit took place.

SURF'S UP

Using an open-source analysis program called Weka J4.8, the researchers wrote scripts to implement their algorithm. By analysing increasing numbers of surfing sessions for each user, they could gradually calculate the minimum number of surfing sessions needed to identify any of the five accurately enough to be of use. After just seven sessions, the algorithm

could identify the users accurately nearly 87 per cent of the time. After 10 sessions, this rose to nearly 92 per cent. By 51 sessions, this was as high as 99.5 per cent. They also found that as the number of users of a site increased, the number of sessions needed to identify a user accurately rose, but not by much.

A further test, taking in 100 users, showed that the number of sessions required for a usable accurate identification (one accurate over 90 per cent of the time) rose to between just 18 and 20. This is still low, especially when you consider that the algorithm used just five general variables as its inputs. More detail should result in greater accuracy from fewer sessions.

There could be a number of practical applications. Supposing, for instance, you use a supermarket's online delivery service to save you the trouble of lugging heavy tins, packets and jars back from the shops. After just a few uses of the service, the site might begin to recognise you and configure itself according to your tastes even as you navigate to the login screen. But one day 'you' go immediately to the vintage champagne and fresh strawberry sections and spend a fortune. Is it you, or someone wining and dining at your expense? By realising that this activity doesn't conform to your clickprint for that site, the online shop may suspect something and alert a customer service representative, who could call you directly before allowing any payment authorisation.

WRITING ANALYSIS

There are less-welcome implications to clickprints, though. Regardless of your browser's privacy settings, online marketers might be able to identify you and bombard you with targeted adverts. This would be done without you logging in or registering, by exploiting your unique site usage profile. Even using a proxy server, which hides your true IP address behind its own, would be useless, because how you surf always gives you away. What's to stop someone impersonating your surfing style? They might then leave messages or even spam in your name on an online forum or as a blog comment.

A team in Arizona says they have a technique that can help prevent that and also ensure that the person you're talking to online really is who they say they are. Ahmed Abbasi and Hsinchun Chen from the Department of Information Systems at the University of Arizona have come up with the concept of 'writeprints'. Similar to clickprints, these also serve as online fingerprints. However, this time a person is identified from entries on blogs, Usenet, online forums and even chatrooms.

CYBER-HOUSE RULES

The pair were concerned that cyberspace allows less-welcome visitors to flourish undetected by passing themselves off as respectable users, often with the support of other 'users', who are in fact the same person using a different online name. "In addition to using the internet as an illegal sales and communication medium, there are several trust-related issues in online communities that have surfaced as a result of online anonymity," said Abbasi. These issues include the problem of accurately identifying strangers you or your children might meet online and with whom you may strike up a friendship. In the real world, you can see who someone is, but online you have only words to



go by. If there were a way to check that person's other posts on other forums, you could spot any anomalies and know all was not what it seemed.

According to Abbasi and Chen, this is now possible. Over time, your writing style expresses itself subtly yet as clearly and uniquely as a fingerprint does to a forensic scientist. Abbasi and Chen have devised a way of analysing your writing style and so identifying the true authors of messages, even when they have used multiple accounts to cover their identities or to vouch for one another. The writeprint identification system created by Abbasi and Chen is grounded in a field of statistics called stylometry. This is the analysis of writing style based on what is written, not to be confused with graphology, which is the analysis of the shapes made when someone puts words on paper.

The idea of identifying the author of a text based on linguistic analysis of other documents they have written is not a new one. Known as forensic linguistics, these techniques have been used in many UK court cases. In the US, an analysis of two texts was used to prove that Theodore Kaczynski was the Unabomber, responsible for killing three and wounding 29 in a protracted campaign of mail bombings.

Authorship analysis systems have included analysing how often someone makes the same grammatical or spelling mistakes, word length and so on. Computerisation of these techniques lends itself readily to software that generates data structures called n-grams, which are words that follow each other in sequence. By comparing the composition and frequency of n-grams, scientists have been able to spot everything from plagiarism in college papers to the authorship of books of the Bible.

MESSAGE PARLOUR

Traditional forensic linguistic analysis produces statistics from any given passage of text. The Arizona team decided to visualise the patterns made by their analyses. Though there have been other attempts to create such visualisations, Abbasi and Chen felt that none was specifically geared to detecting online deception.

The basis for their new approach is principal component analysis; see below for details. A web spider accesses and stores in a large database all the messages left on a number of

public web forums. Feature-extraction software analyses the captured messages for a rich set of features. This includes parts of speech such as adjectives and adverbs, plus more statistical concepts such as sentence complexity.

Next, a program running a sliding window algorithm generates more data based on the data extracted already, like a biologist culturing bacteria from a tiny sample to see it more clearly. The result of this amplification process is the writeprint itself. The system of x and y coordinates for the generated data are calculated. These coordinates are displayed and an organic-looking picture emerges. Onscreen, the software splits the writeprint into six parts, each containing a unique shape. These in turn cover the distribution of word lengths, the use of punctuation and special characters, sentence structure, letter frequencies and even jargon.

When presented with a sample of text that is supposed to have been written by a particular person, calling up his writeprint and taking the writeprint of the text sample could identify him more easily than a fingerprint test and with a great deal of accuracy. When given an anonymous message, it's easy to generate a writeprint for it and search the database for a similar one.

When presented with just one example of a message that's known to have come from a user, writeprints proved less effective at recognising other texts from the same author than a computer running support vector machine (SVM) pattern-recognition software. However, when given 10 or more such messages, writeprints were 100 per cent accurate, where the SVM system runs at around 90 per cent.

As with the other biometric profiling ideas, however, there are still problems to resolve. "[Writeprints] are constrained when dealing with shorter individual messages; that is, messages of fewer than 30 to 40 words," said Abbasi. "This is due to the minimum length needs of the sliding window algorithm." Some common written communications such as text messages are too short for sliding window algorithms to amplify accurately enough.

CALM BEFORE THE STORM

If they could be made more accurate, the applications for calm identification techniques

such as these are potentially wide-ranging. In a conversation, you don't usually stand up, announce who you are and begin to say your piece. The interplay is far more dynamic and informal. Writeprints could facilitate interactions between individuals that are just as natural by giving us the ability to recognise each other from what we type and also spot impostors regardless of their usernames.

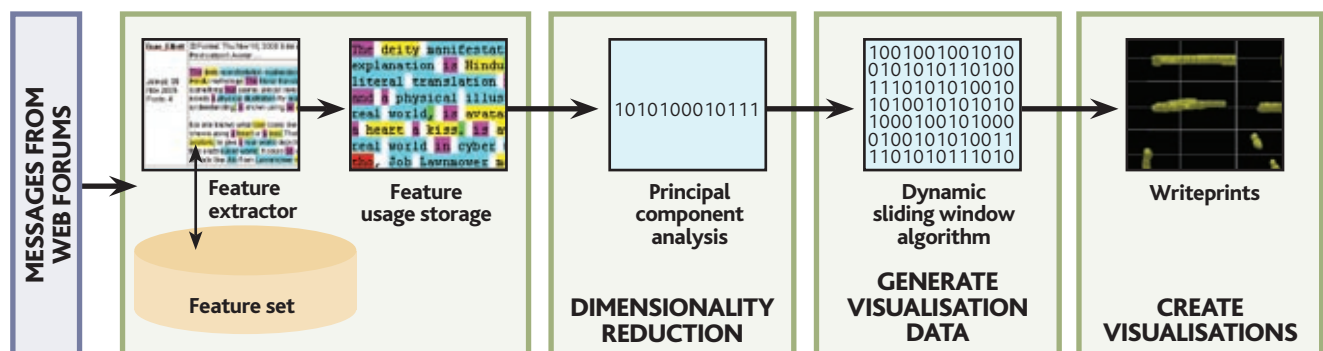
"The use of [writeprints] specifically tailored towards multilingual online messages and automatic recognition mechanisms makes our proposed visualisations feasible for identification of online messages," said Abbasi. The implication is that writeprints work regardless of the language used. They should work as well using Arabic as they do English. Being able to identify writeprints automatically in any language may help prevent online crime. However, there might be serious consequences for internet users in other countries.

While calm-computing analysis might serve to protect the individual, there's always the possibility that these techniques could end up being used as tools of repression. In an authoritarian state, being able to identify and track a single user out of millions, regardless of their attempts to defy detection, could be used to the detriment of personal freedom. Maybe the most sinister aspect of these techniques is that, by being 'calm', we may never even realise they're taking place. Your mobile phone might one day inform on who's carrying it, as well as its location, as a matter of course.

IDENTITY PARADE

For now, we have nothing to fear from these advances. If they do come into general use, they could be far more accurate than current biometric identification techniques. These currently all have lower accuracies than most people suspect. In a 2004 Home Office report on the user experience of biometric scanning, the figures for successful identifications were surprisingly low. While iris scans have an upper success rate of 96 per cent, this falls to 81 per cent for fingerprint scanning and just 69 per cent for facial recognition. A calm alternative could do away with all three, replacing them with 100 per cent accuracy, all without you ever knowing they were being used. **ES**

Identify yourself How the writeprints algorithm works



▲ First, messages collected from web forums have a set of features extracted. These can be anything, including word lengths or even sentence complexity, depending on what the system has had programmed into its feature set, so the algorithm can work in any language. The features of the messages are then stored and passed on to the principal component analysis

process. This performs a function called dimensionality reduction, in which it analyses the number of times each feature occurs in the data being analysed. The output is a faint but distinct and unique pattern of data. The dynamic sliding window algorithm then generates more data based on what it already has, which can then be visualised graphically as a writeprint.